CUSTOMER SEGMENTATION ANALYSIS REPORT

Abstract

This report demonstrates an optimized K-means clustering approach for customer segmentation, leveraging key features. The resulting visualizations provides actionable insights for targeted business strategies.

Emuejevoke Eshemitan

vokeeshemitan@gmail.com

1. Introduction

This report presents a comprehensive analysis of customer segmentation using clustering techniques. The objective is to group customers based on their purchasing behavior, enabling businesses to tailor marketing strategies and services to distinct customer segments. The dataset used for this analysis comprises various features related to customer transactions.

Dataset Overview:

- The dataset contains 3900 entries.
- There are 18 columns in the dataset.
- No missing values are present in any of the columns, as the non-null counts for each column are equal to the total number of entries (3900).
- The dataset includes integer (int64), float (float64), and object (categorical) data types.
- Numerical features: 'Age', 'Purchase Amount (USD)', 'Review Rating', 'Previous Purchases'.
- Categorical features: 'Gender', 'Item Purchased', 'Category', 'Location', 'Size', 'Color', 'Season', 'Subscription Status', 'Shipping Type', 'Discount Applied', 'Promo Code Used', 'Payment Method', 'Frequency of Purchases'.

2. Data Exploration and Preprocessing

The first step involved exploring the dataset to understand its structure and characteristics. The dataset was examined for missing values, and fortunately, no such issues were identified. The features were categorized into numerical and categorical variables. Summary statistics, such as mean and standard deviation, were computed for numerical features to gain insights into the data distribution.

Findings From Data Exploration:

Customer demographics:

- Age spans 18 to 70, with an average age of 44 years.
- Purchases range from $20 to $100, averaging around $59.76.
- Review ratings range from 2.5 to 5.0, with an average rating of 3.75.
- Previous purchase frequency ranges from 1 to 50, averaging about 25.35.

Purchase details:

- Gender distribution: Two values (Male, Female), with Male being most common (2652 occurrences).

- Items purchased: 25 unique items, with 'Blouse' as the most frequent (171 occurrences).
- Categories: 4 unique categories, with 'Clothing' being the most prevalent (1737 occurrences).
- Locations: 50 unique locations, with 'Montana' being the most frequent (96 occurrences).
- Sizes: 4 unique sizes, with 'M' being the most prevalent (1755 occurrences).
- Colors: 25 unique colors, with 'Olive' being the most common (177 occurrences).
- Seasons: 4 unique seasons, with 'Spring' being the most frequent (999 occurrences).

Transaction and engagement specifics:

- Subscription status: Two values ('Yes', 'No'), with 'No' being the most common (2847 occurrences).
- Shipping types: 6 unique types, with 'Free Shipping' being the most frequent (675 occurrences).
- Discounts applied: Two values ('Yes', 'No'), with 'No' being the most common (2223 occurrences).
- Promo codes used: Two values ('Yes', 'No'), with 'No' being the most common (2223 occurrences).
- Payment methods: 6 unique methods, with 'PayPal' being the most frequent (677 occurrences).
- Frequency of purchases: 7 unique frequencies, with 'Every 3 Months' being the most common (584 occurrences).

3. Feature Engineering

3.1 Categorical Variable Transformation

Categorical variables were transformed using label encoding to convert them into a format suitable for clustering algorithms. This step was essential to ensure that the algorithm could effectively process and interpret these variables during the clustering process.

3.2 Feature Scaling

To ensure that all features contribute equally to the clustering process, numerical features were scaled using the StandardScaler. Scaling prevents features with larger scales from dominating the clustering algorithm and ensures a more balanced contribution from all features.

4. Clustering Methodology

The K-Means clustering algorithm was chosen for this analysis due to its simplicity, efficiency, and effectiveness in identifying compact, spherical clusters. The primary goal was to determine the optimal number of clusters (K) that best represent the underlying patterns in the data.

4.1 Determining Optimal K

The Elbow Method was employed to identify the optimal number of clusters. This involved running the K-Means algorithm for a range of K values and plotting the within-cluster sum of squares (WCSS) against the number of clusters. The 'elbow' in the plot represents the point of diminishing returns, suggesting the optimal number of clusters.

4.2 K-Means Modeling

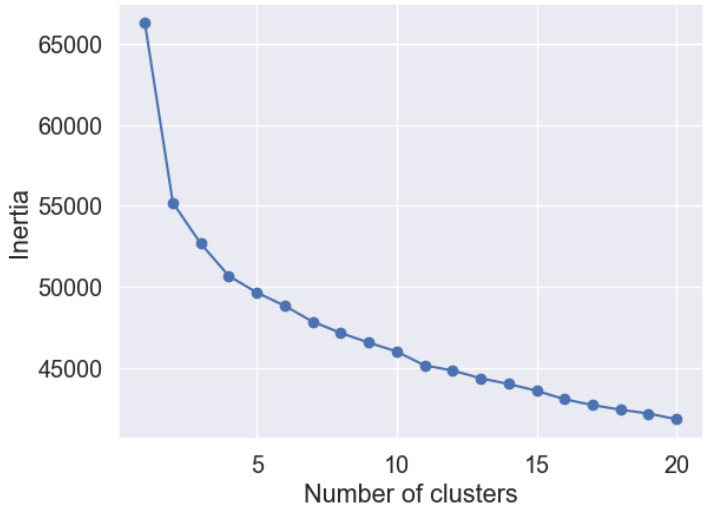


Figure 1: Elbow Plot for determining optimal number of clusters

Based on the Elbow Method, K=4 was selected as the optimal number of clusters. K-Means models were built for K=4, K=5, and K=6 to explore alternative cluster configurations.

5. Evaluation Metrics

Two primary metrics, Inertia (within-cluster sum of squares) and Silhouette Score, were used to evaluate the quality of clustering.

5.1 Inertia

Inertia measures the compactness of clusters. A lower inertia indicates tighter clusters. The Inertia values after K=4 exhibited low change in inertia, suggesting better-defined clusters.

5.2 Silhouette Score

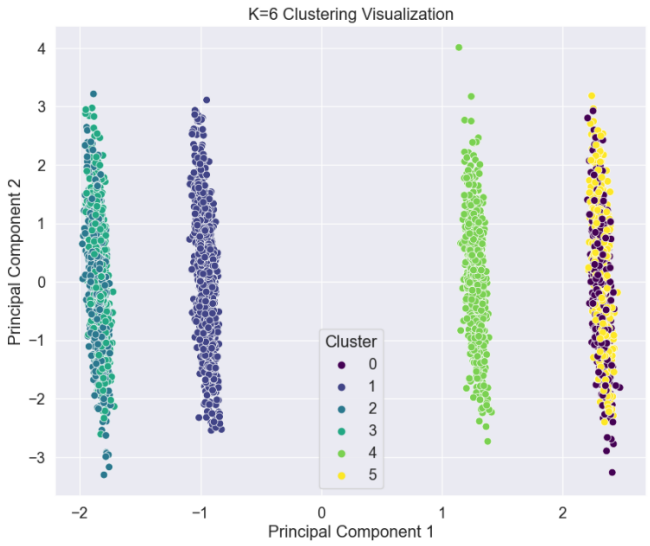
The Silhouette Score measures the separation between clusters. Scores range from -1 to 1, with higher values indicating better-defined clusters. Silhouette Scores were computed for K=4, K=5, and K=6, and K=4 consistently outperformed the other configurations.

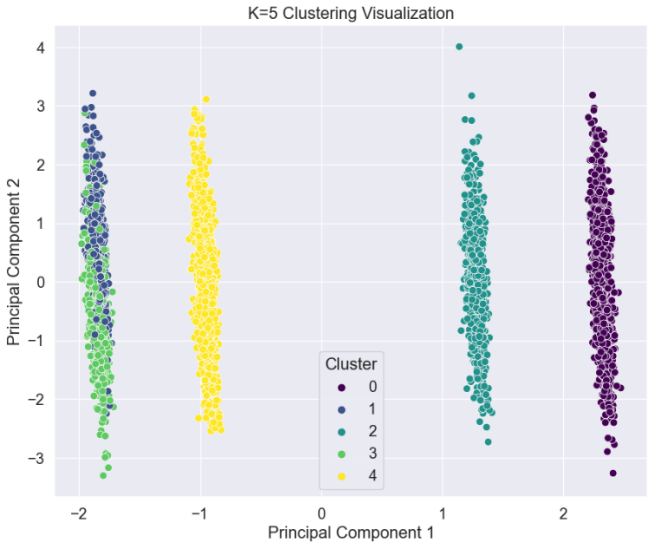
6. Insights Derived from Customer Segments

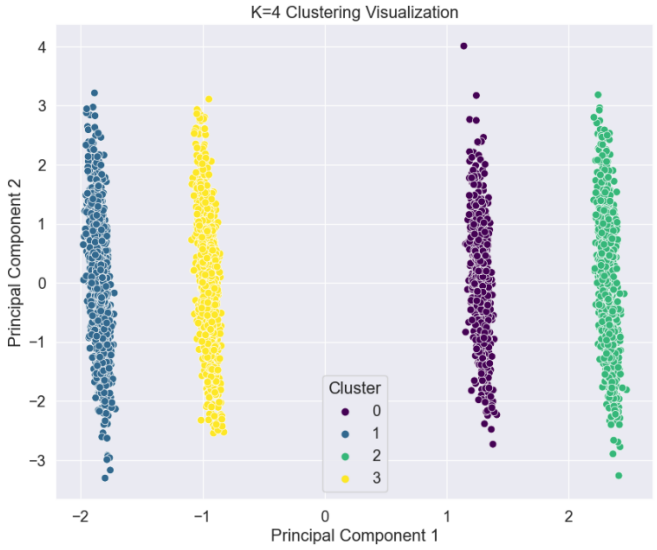
6.1 Visualization

To gain a visual understanding of the clusters, principal component analysis (PCA) was applied to reduce the dimensionality of the data. 2D and 3D visualizations were created for K=4, K=5, and K=6.

2D Visualization







3D Visualization

To interact with the 3D visualization, refer to the Jupiter notebook.

6.2 Customer Segment Characteristics

This section contains characteristics and potential customer segments for each cluster:

Table 1: Customer Segment Characteristics

Cluster	Gender	Category Preference	Subscription Status	Discount Applied	Promo Code Used	Potential Customer Segment
0	Mostly Male	Clothing, Outerwear, and Footwear	No specific pattern observed	No	Yes	Fashion-forward males who prefer a variety of clothing items and are likely to use promotional codes.
1	Mostly Female	Clothing, Outerwear, Accessories, Footwear	No specific pattern observed	No	No	Female customers who purchase a variety of fashion items without using discounts or promo codes.
2	Mostly Male	Clothing, Footwear, Accessories	Yes	Yes	Yes	Male customers who are subscribed, frequently use discounts and promo codes, and have a preference for clothing, footwear, and accessories.
3	Mostly Male	Clothing, Footwear, Accessories	No	No	No	Non-subscribed male customers who do not frequently use discounts or promo codes but still purchase clothing, footwear, and accessories.

7. Optimization and Final Results

7.1 Optimization Approach

To enhance the clustering optimization process, an iterative feature selection approach was implemented. The initial set of features considered for clustering included various customer attributes such as 'Age,' 'Gender,' 'Item

Purchased,' and more. Three iterations were conducted, progressively refining the feature set to identify the most impactful attributes for optimal clustering.

7.2 Iterative Feature Selection

Iteration 1: Resulted in a Silhouette Score of 0.0861.

Iteration 2: Resulted in a slightly improved Silhouette Score of 0.0879.

Iteration 3: Resulted in a significantly improved Silhouette Score of 0.3976, surpassing the initial scores.

7.3 Results from Optimization

After several iterations, it was determined that utilizing the features 'Gender', 'Item Purchased,' 'Category,' 'Subscription Status,' 'Discount Applied,' and 'Promo Code Used' led to a substantially higher Silhouette Score for k=4, reaching 0.3976 compared to the initial score of 0.0861.

8. Conclusion

The analysis revealed that K=4 is the most suitable configuration for customer segmentation based on both Inertia and Silhouette Score. The identified customer segments offer valuable insights into their behaviors and preferences, enabling businesses to tailor marketing efforts, improve customer satisfaction, and optimize product offerings.